Klassensitzungen

Vorhersagen von Struktur-Aktivitäts-Beziehungen für die Wirkstoffforschung*

Knut Baumann

Institut für Medizinische und Pharmazeutische Chemie, TU Braunschweig
Beethovenstraße 55, D-38106 Braunschweig

Wirkstoffe durchlaufen einen aufwändigen Optimierungsprozess von ihrem ersten Auffinden bis hin zum fertigen Arzneistoff. In vielen Optimierungsschritten wird der Wirkstoff abgewandelt, um eine ganze Schar von wichtigen Eigenschaften zu verbessern. Dieser Prozess ist sehr langwierig und teuer. Die Auswirkung einer chemischen Modifikation des Wirkstoffs auf die untersuchte Eigenschaft ist dabei vielfach nicht bekannt. Wäre sie bekannt, dann könnte die Optimierung deutlich zielgerichteter betrieben werden. Hier setzt die Analyse von Struktur-Aktivitäts-Beziehungen an. Deren Ziel ist es, eine qualitative oder quantitative Beziehung zwischen der Bioaktivität (oder ganz allgemein einer bestimmten Eigenschaft) eines Wirkstoffs und dessen chemischer Struktur herzustellen. Bereits im Jahr 1868 postulierten die schottischen Pharmakologen A. Crum-Brown und T.R. Frazer, dass die physiologische Wirkung (Φ) einer Substanz eine Funktion der chemischen Konstitution (C) sein müsse [1]. Sie benutzen dabei die folgende Gleichung, um den Sachverhalt mathematisch darzustellen:

$$\Phi = f(\mathrm{C}).$$

Nur in Ausnahmefällen gelingt es die Funktion f in dieser Gleichung zu bestimmen. Wesentlich einfacher ist es die Veränderung der physiologischen Wirkung ($\Delta\Phi$) in Abhängigkeit von der Veränderung der chemischen Konstitution (ΔC) zu modellieren:

$$\Delta\Phi = g(\Delta\mathrm{C}).$$

Die besondere Herausforderung ist es dabei die chemische Konstitution so zu beschreiben, dass sich damit ein bedeutungsvolles mathematisches Modell erstellen lässt. Chemische Strukturformeln sind sehr gut geeignet chemische Moleküle und deren Reaktionen zu beschreiben. Sie sind jedoch nicht ohne weiteres als Eingabe für die Funktion g geeignet. Dazu muss die chemische

---

* Der Vortrag wurde am 14.02.2014 in der Klasse für Mathematik und Naturwissenschaften der Braunschweigischen Wissenschaftlichen Gesellschaft gehalten.

Information zunächst in einen Zahlenstrang kodiert, d.h. übersetzt werden. Mit dieser numerischen Repräsentation steht und fällt die Analyse. Ist sie ungeeignet, wird sich für die Funktion *g* keine geeignete Form finden lassen. Mittlerweile steht für die numerische Repräsentation chemischer Moleküle zur Analyse der Struktur-Aktivitäts-Beziehungen eine Vielzahl an Möglichkeiten zur Verfügung [1]. Keine ist perfekt, sie muss dem Problem jeweils angepasst werden. Vielfältig einsetzbar sind sogenannte chemische Fingerabdrücke. Bei Letzteren handelt es sich um einen Vektor pro Molekül. Jeder Eintrag in diesem Vektor gibt die Anzahl einer bestimmten chemischen Substruktur im Molekül an. Welche Substrukturen erfasst werden ist in einer Bibliothek an Substrukturen festgelegt oder kann auch durch einen Satz an Rechenvorschriften beschrieben werden. Nachdem die Moleküle kodiert sind und deren biologische Eigenschaften experimentell ermittelt wurden, werden Techniken des Maschinellen Lernens angewendet, um die Funktion *g* zu ermitteln. Mit Hilfe eines Trainingsdatensatzes, der aus der numerischen Repräsentation einer Schar an Molekülen und deren zugehöriger biologischer Aktivität besteht, wird die Funktion *g* erlernt. Im einfachsten Fall kann es sich um ein lineares Modell handeln. Bei schwieriger zu modellierenden Eigenschaften werden nichtlineare Modelle wie Wälder von Entscheidungsbäumen (sog. Random Forests) eingesetzt [2].

Zwei Eigenschaften dieser Modelle sind dabei von großer Bedeutung. Zum einen sollen die Modelle für Medizinische Chemiker intuitiv zu erfassen sein, so dass der nächste Optimierungsschritt aus dem Modell leicht abzuleiten ist. In dieser Hinsicht sind lineare Modelle ideal. Hier kann der Effekt einer Veränderung der chemischen Struktur sehr leicht ermittelt und auch visualisiert werden. Im Gegensatz dazu ist das bei Random Forests nicht einfach möglich, selbst dann, wenn die Funktion *g* sehr gut gelernt wurde. Zum anderen müssen die Modelle möglichst zuverlässige Vorhersagen liefern, damit der Optimierungsprozess in die richtige Richtung geleitet wird. Um die Zuverlässigkeit der Vorhersagen beurteilen zu können, wird mit dem oben erwähnten Trainingsdatensatz der Vorhersagefehler des Modells mit Hilfe der Kreuzvalidierung oder des Bootstrappings bestimmt [3]. Es zeigt sich, dass je nach chemischer Klasse des Eingabemoleküls der Vorhersagefehler im modellierten Datenraum stark variieren kann. Ideal wäre folglich die Angabe eines individuellen Zuverlässigkeitsmaßes (im Sinne eines Vorhersageintervalls) für jedes Molekül statt eines globalen Vorhersagefehlers für das gesamte Modell. Derartige Techniken werden derzeit intensiv in der Statistik erforscht [4,5] und in weiterer Folge auf ihre Anwendbarkeit für die Analyse der Struktur-Aktivitäts-Beziehungen überprüft werden.

Neben der quantitativen Modellierung des Zusammenhangs zwischen Struktur und Aktivität, gewinnt die Visualisierung großer chemischer Datensätze (> 1000 Moleküle) zunehmend an Bedeutung. Durch öffentlich zugängliche Datenbanken (z.B. ChEMBL [https://www.ebi.ac.uk/chembl/]) oder in den jeweiligen Insti-

tutionen erstellte Datenbanken stehen Medizinischen Chemikern heute riesige Datenbestände zur Verfügung, die durch manuelles Inspizieren der Daten nicht mehr effizient auszuwerten sind. Ein automatisiertes Gruppieren und Anordnen der Moleküle nach Stoffklassen und biologischer Aktivität ist hier von Nöten. Bevor diese Gruppierung vorgenommen werden kann, müssen die Moleküle erneut kodiert werden. Würde man die Gruppierung und Anordnung auf Basis von chemischen Substrukturen vornehmen, so wie es oben bei den chemischen Fingerabdrücken beschrieben wurde, dann würde eine viel zu kleinteilige Gruppierung resultieren und die Visualisierung dieser Gruppierungen würde sehr unübersichtlich werden, so dass kein Informationsgewinn für den Medizinischen Chemiker resultiert. Folglich muss die chemische Information der Moleküle stark abstrahiert werden. Das gelingt dadurch, dass ganze Substrukturen eines Moleküls zu Pseudoatomen zusammengefasst werden. Die Pseudoatome werden so ausgewählt, dass sie das Potential für nicht-kovalente Interaktionen des zu kodierenden Wirkstoffs mit Rezeptoren oder Enzymen widerspiegeln (eine chemische Reaktion des Wirkstoffs mit Enzymen oder Rezeptoren stellt die Ausnahme dar und wird deshalb nur in Spezialfällen kodiert). Typischerweise werden elektrostatische Interaktionen, das Potential für Wasserstoffbrückenbindungen und das Potential für sogenannte hydrophobe Wechselwirkungen kodiert. Positiv oder negativ ionisierbare funktionelle Gruppen (PI/NI) können beispielsweise mit positiv oder negativ geladenen Aminosäuren aus Rezeptorproteinen interagieren. Zur Abstraktion der Moleküle wird nun jede negativ ionisierbare Gruppe im Wirkstoff gegen das Pseudoatom NI ersetzt, unabhängig von der exakten chemischen Struktur der jeweiligen Gruppe. Die chemischen Bindungen zu den Nachbargruppen bleiben dabei erhalten. Genauso wird mit positiv ionisierbaren Gruppen und allen anderen Interaktionstypen verfahren. Auf diese Weise entstehen Pseudomoleküle, die auf das Interaktionspotential beschränkt sind und wesentlich weniger komplex sind als die „echten“ Wirkstoffe. Durch diese Reduktion der Moleküle auf das Wesentliche können komplexere Rechnungen durchgeführt und größere Datensätze verarbeitet werden. Zur Anordnung und Gruppierung der Moleküle eines großen Datensatzes werden nach der Kodierung die Pseudomoleküle verglichen. Immer dann, wenn zwei Pseudomoleküle in bestimmen Pseudomolekülteilen übereinstimmen werden sie in einem hierarchischen Netzwerk einer Gruppe zugeordnet. Je größer die Übereinstimmung ist, desto höher werden die Moleküle in der Hierarchie angeordnet. Auf diese Weise entsteht ein Netzwerk, welches auf der höchsten Hierarchieebene Gruppen von Molekülen mit sehr ähnlichem Interaktionspotential anzeigt. Die Analyse der Moleküle in den Gruppen und deren Beziehungen in dem Netzwerk erlauben es dem Medizinischen Chemiker schnell in großen Datensätzen zu navigieren und Struktur-Aktivitäts-Beziehungen abzuleiten [6]. Das Erstellen des Netzwerkes erfolgt lediglich durch Analyse der Ähnlichkeit der Pseudomoleküle. Färbt man allerdings das Netzwerk mit den biologischen Aktivitäten der gruppierten Moleküle ein, so stellt man fest, dass

sich sehr homogene Gruppen bezüglich der Bioaktivität in diesem Netz bilden. Das zeigt, dass die gewählte Molekülrepräsentation auf Basis des Interaktionsmusters sehr gut geeignet ist.

Kern des Verständnisses der Variation der Bioaktivität verschiedener Moleküle an einer bestimmten biologischen Zielstruktur ist das Wissen darüber, wie die chemische Struktur die biologische Aktivität beeinflusst. Mit Techniken der Chemieinformatik lässt sich diese Information wie oben beschrieben vielfach sehr effizient extrahieren.

Literatur

[1] a)Baumann, K. 1999: Uniform-length molecular descriptors for Quantitative Structure-Property Relationships (QSPR), Quantitative Structure-Activity Relationships (QSAR), classification studies, and similarity searching. TrAC **18**: 36–46. b) Stiefl, N. & K. Baumann 2003: Mapping Property distributions of molecular surfaces (MaP): Algorithm and evaluation of a novel 3D Quantitative Structure-Activity Relationship technique. J. Med. Chem. **46**: 1390–1407.

[2] Breiman, L. 2001: Random Forests. Mach. Learning **45**: 5–32.

[3] Baumann, K. 2003: Cross-validation as the objective function for variable selection techniques. TrAC **22**: 395–406.

[4] Politis, D.N. 2013: Model-free model-fitting and predictive distributions. Test **22**: 183–221.

[5] Efron, B. 2014: Estimation and accuracy after model selection. J. Amer. Statist. Assoc. **109**: 991–1007.

[6] Wollenhaupt, S. & K. Baumann 2014: InSARa: Intuitive and interactive SAR interpretation by reduced graphs and hierarchical MCS-based network navigation. J. Chem. Inf. Model. **54**: 1578–1595.